Cross Lingual Information Retrieval Using Search Engine and Data Mining

Mallamma V Reddy[1], Dr. M. Hanumanthappa[2], Manish Kumar[3]
[1,2]Department of Computer Science and Applications,
Bangalore University, Bangalore, INDIA
[1]mallamma_vreddy@yahoo.co.in
[2]hanu6572@hotmail.com
[3]Department of Master of Computer Applications,
M. S. Ramaiah Institute of Technology, Bangalore-560 054, INDIA
[3]manishkumarjsr@yahoo.com

Abstract:- **With the explosive growth of international users, distributed information and the number of linguistic resources, accessible throughout the World Wide Web, information retrieval has become crucial for users to find, retrieve and understand relevant information, in any language and form. Cross- Language Information Retrieval (CLIR) is a subfield of Information Retrieval which provides a query in one language and searches document collections in one or many languages but it also has a specific meaning of cross-language information retrieval where a document collection is multilingual. In the present research, we focus on query translation, disambiguation of multiple translation candidates and query expansion with various combinations, in order to improve the effectiveness of retrieval. Extracting, selecting and adding terms that emphasize query concepts are performed using expansion techniques such as, pseudo-relevance feedback, domain-based feedback and thesaurus-based expansion. A method for information retrieval for a query expressed in a native language is presented in this paper. It uses insights from data mining and intelligent search for formulating the query and parsing the results.**

Keywords: **Cross Lingual Information Retrieval, Heuristic Method, Text Categorization**

I. INTRODUCTION

Cross-Language Information Retrieval (CLIR) is where the user request and the document collection against which the request is to be matched are in two different human languages. The aim of CLIR is to match the request against the collection as if the request had been issued in the document collection language to begin with. This kind of system is useful in the situation where a user who can read several different languages wants to find information in a collection containing documents in many languages, while avoiding the work involved in formulating multiple requests. Cross-language information retrieval [1] enables users to enter queries in languages they are fluent in, and uses language translation methods to retrieve documents originally written in other languages. The scope for Cross-Language Information Access [2] goes beyond the Cross-Lingual Information Retrieval (CLIR) paradigm by incorporating query disambiguation as well as post search processing. The key emphasis is on the relevance of the results. The cross lingual information access paradigm may take the form of machine translation of snippets, summarization and subsequent translation of summaries and/or information extraction from the target language. CLIR has three basic approaches [3] : a) document translation - where the queries are posed to existing document repositories, b) query translation - where the queries are translated into the target language and results displayed, and c) inter lingual translations - where queries and results are translated. Our approach is a variation of the third category.

Objectives of our research work are to:-

- Develop an approach for cross lingual information retrieval for queries expressed in the native language.
- Use data mining techniques to cluster the results and retrieve a resultant set closest to the user's query, and
- Present the results in various display methods to the user.

The key aspect of the proposed approach is as follows. It is composed of two distinct aspects: preprocessing the query to identify the query's meaning and post-processing the results for relevance match. In the preprocessing stage, the query will be expanded and the expanded query is presented to the search engine. In the post processing stage, based on the relevance match of the retrieved content, the resultant documents will be reordered and presented to the user. The initial feedback from the users seems to indicate that the relevance of the retrieved documents is higher than the conventional approach. However, the time needed to perform the processing is a significant factor.

Introduction to Information Retrieval

An Information Retrieval System is defined as any system that matches a user request against a document collection, returning a list of documents considered relevant to the request. The user request is an expression of a user information need. For example, the user might issue the following request. Have you got any documents pertaining to the Clinton Lewinsky scandal, particularly regarding his testimony before Congress?

An automatic IR system then usually carries out some processing on the user request to derive a form of the request that it can match directly against the document collection using some form of matching algorithm. The processed request, which may take many forms, is known as the query. . Query formats commonly employed in the IR world include the natural language query, where the request is not processed much at all, and the bag of words format, where

DOI: 01.IJIT.01.02.114

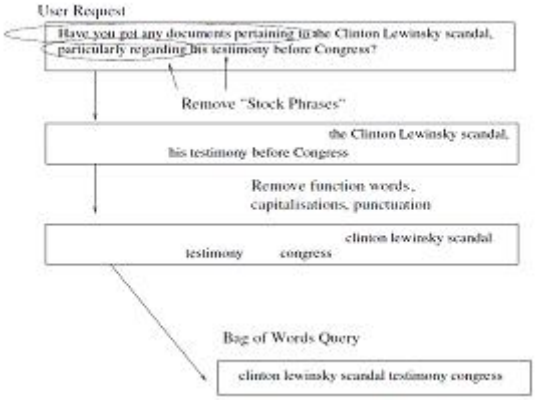

Figure 1: Processing a Request to Extract a Key Words

function words, punctuation and phrases like "on the subject of" are removed from the request, suffixes stripped, and a selection of what are known as keywords extracted to form the query "Fig. 1". For example, the user request above, when processed in this manner, would yield the bag of words query: clinton lewinsky scandal testimony congress Commonly used search engines, such as Google ask the user to enter the bag of words query directly, thereby circumventing part of the request processing stage by getting the user to do some preprocessing in her head. The document collection consists of a set of individual documents, each of which identifies a single text, such as a book, journal article, or web page. The documents in the collection can consist of the texts themselves, such as, for example, the document database of a web search engine, or of summaries that point the user toward the texts, such as book titles in a conventional electronic library catalogue.

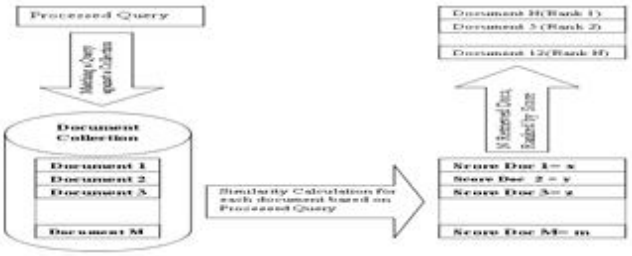

Figure 2:- Document Scoring and Extraction Based on User Query

The former case is known as full text retrieval and is the type of document collection we are interested in our experiments. The document collection is also usually processed in an identical manner to user requests when it is being compiled. The query is matched against the document collection using a matching algorithm which calculates a score for each document in the collection reacting its perceived similarity to the query Fig 2. Similarity scores may be based simply on the frequency of individual query terms (words or phrases), or may exploit term weights (scores per term) calculated using frequency data. The Vector Space Method and the Probabilistic Model of Information Retrieval (which is the model employed by the IR system used in the experiments discussed in subsequent chapters) provide well-founded ways of doing this. Generally, a list of the N most closely matching documents is returned to the user. This list of returned documents is often called the retrieved document list. The aim is to retrieve as many relevant documents as possible in this list, while avoiding the retrieval of irrelevant ones. IR systems are generally evaluated using two metrics, precision and recall. Precision is defined as the proportion of retrieved documents which are actually relevant to the query derived from the user request.

Precision = Not Relevant / Total Retrieved

Recall is the proportion of documents known to be relevant to the query in the entire collection that have been retrieved in the retrieved document list for that query.

Recall = Not Relevant Retrieved / Total Known Relevant

II. PROPOSED APPROACH

The proposed approach "Fig. 3" is composed of two distinct and complementary stages, namely, preprocessing and post processing. In the preprocessing stage, the search query in an Indian language is parsed and disambiguated using the lexicons available for that Indian language and the initial search terms are translated into English. These English terms may be further disambiguated using WordNet and other ontologies and the expanded query is submitted to the search engine. In the post processing stage, the results from the search engine are summarized, mapped to the target language and the results presented to the user. The results in English need to be summarized for relevance and displayed to the user.

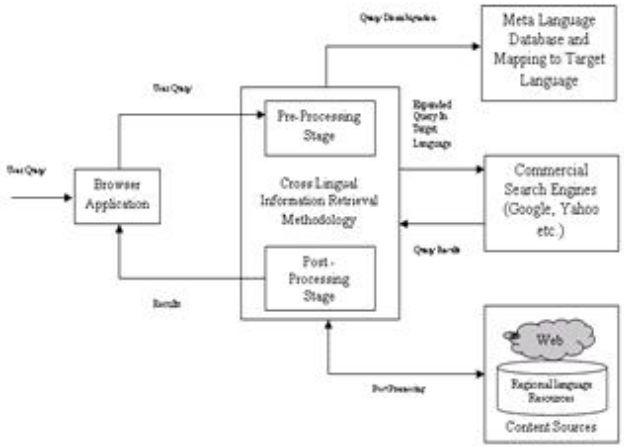

Figure 3:- Complete Process Architecture

Preprocessing stage: The preprocessing stage is composed of four distinct steps. These steps have been adapted from the work of Storey [6].

Step 1 - Query Parsing: It involves parsing the natural language query specified by the user in regional language. First the query is segmented and the words are disambiguated using an ontology and/or other lexicons available in that language. Then the initial query is translated into English.

Step 2 – Query Expansion: The output of the parsing step is a set of initial translated query terms which become the input to the query expansion step. The query expansion process involves expanding the initial query using lexicons and ontologies. It also includes adding appropriate personal information as well as contextual information relevant to the query. Since ontologies contain domain specific concepts,

DOI: 01.IJIT.01.02.114

appropriate hypernym(s) and hyponym(s) are added as mandatory terms to improve precision.

Step 3 – Query Formulation: The expanded set becomes the input to the query formulation step. In this step, the query is formulated according to the syntax of the search engine. For each query term, the synonym, Hypernym and hyponym and the negative knowledge is added with an OR, AND & NOT operator.

Step 4 – Search Knowledge Sources: This step submits the query to one or more search engines (in their required syntax) for processing. The query construction heuristics can work with most search engines.

Post processing stage

In this stage, the results from the search engine (URLs and 'snippets' provided from the web pages) are retrieved, and translated to the target language. Available lexicons and ontologies are also used in the translation. Further, a heuristic result processing mechanism described below is used to identify the relevance of results retrieved with respect to the source language. Finally, these are aggregated and the resultant summarized content is presented to the user. The approach used is from the information retrieval [4] perspective, which also integrates the insights gained from data mining. Our approach visualizes the problem of categorizing the results akin to the results merging approach [5]. The objective of the post processing is to aggregate the results for the given query, rank and reorder the results, and present the results in a new sorted order based not only on the output of the search engine but also on the content of the retrieved documents. Thus, these downloaded documents are indexed and comparable document scores for the downloaded documents are calculated. The metrics utilized are word relevance, word to document relationship and clustering strategies. Finally, all the returned documents are sorted into a single ranked list along with display mechanisms for helping the user view and decide on the results. These measures have different connotations in traditional data mining .

a) Candidate word: words that have high information gain in a given document.
b) Information gain: the parametric importance of a word in the retrieved document.
c) Pair-wise measure: the correlations between the candidate word and the input keywords are found using the pair wise measure. They also give the relationship between words and help disambiguate similar words.
d) Cluster relationship: this gives a measure of the degree of clustering of candidate words in a given document. The post processing stage of the proposed information retrieval method has the following five distinct steps:

A) Parameter estimation:

This step scans the document and identifies at a high level the various keywords that are present. Two methods have been applied in the pre-processing stage 1) removal of stop words (stop word list) and 2) stemming algorithm: porter's stemming algorithm. After the pre-processing,

i) The information gain of keywords is calculated using statistical methods. Information gain is defined as the number of times the word (meaning) occurs in the document.
ii) For words with maximum information gain, the pair wise relationship of each keyword is also observed.
iii) The mutual correlation between each word and all the other candidate words are found. This is continued till the features with maximum correlations are identified and sorted. The cluster relationship is the measure of distribution of the word in the document.

B) Categorization:

The words, which satisfy the threshold limits for each of the three parameters, are selected as candidate words. These three measures help establish relationships between the words and give a measure of the relationships in the documents. The document map (intermediate structure) is an intermediate representation displaying the key candidate words at the place of occurrence. The document maps help in characterization, differentiating and grouping content.

C) Aggregation:

The same method (steps *a* and *b*) is applied for all the documents and the results are aggregated. Based on this, the documents "Fig. 4" are re-ordered using the parametric results and the categorization stage. This stage eliminates irrelevant documents, aggregates different versions of the same document and organizes the overall display to be of documents that are relevant and closest to the user query.

D) Display processing:

At this stage, along with the document map, a word-by-word translated map in user's language is also shown. Thus the users can select documents that they feel are relevant from the map in their own language.

E) Learning:

Based on the results of the above stages, the learning algorithm assigns weights to the parameters and learns (using machine learning methods) from the selection options made by the users. The feedback and suggestions from the users are collected for document mapping. Simultaneously, the above method is applied for documents in the native language where parallel corpora are available.

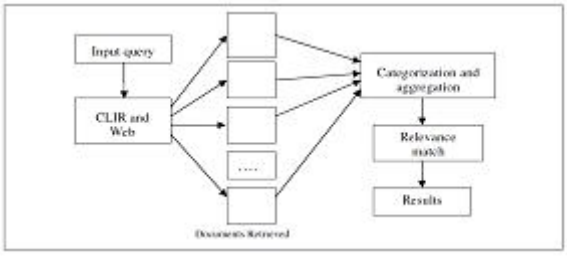

Figure 4:- Results Transformation Method

The key aspects in this system are the mapping mechanism between words in different languages, aggregation method at document/repository level, structure of document maps and the learning system. For each candidate word the pair wise measure gives a measure of correlation. However, these

DOI: 01.IJIT.01.02.114

correlations are not available in dictionary representations and must be generated by use of appropriate ontological systems. Thus, a search and traversal method that navigates the ontology for distance measure is crucial for finding the pair wise and relationship measures.

CONCLUSION

This paper has outlined an approach for Cross Lingual Information Retrieval which emphasizes pre and post processing strategies for the queries entered in a source language. Mechanisms for displaying the results have been outlined which give the users a better idea of what the documents contain. The approach works on top of existing search engines and helps refine the search process further.

REFERENCES

1. Ballesteros, L. and Croft, W. B. (1997) Phrasal Translation and Query Expansion Techniques for Cross-Language, Information Retrieval, *Proceedings of SIGIR'97*.

2. CLIA Research Report, "Development of Cross Lingual Information Access (CLIA) System" funded by Government of India, Ministry of Communications & Information Technology, Department of Information Technology (No. 14(5)/2006 – HCC (TDIL) Dated 29-08-2006) retrieved from www.mt- archive.info/ IJCNLP-2008-CLIA.pdf on Feb 21, 2009.

3. Maeda, A., Sadat, F., Yoshikawa, M., and Uemura, S. (2000) Query term disambiguation for Web cross- language information retrieval using a search engine, *Proceedings of the fifth international workshop on Information retrieval with Asian languages*, pp.25-32, September 30-October 01, Hong Kong, China.

4. R. Shriram, Vijayan Sugumaran, AMCIS 2009 Proceedings Americas Conference on Information Systems (AMCIS) Cross Lingual Information Retrieval Using Data Mining Methods.

5. Si, L., Callan, J., Cetintas, S., Yuan, H. (2008) An effective and efficient results merging strategy for multilingual information retrieval in federated search environments, *Information Retrieval*, Vol. 11 , No. 1, February, pp. 1 – 24.

6. Storey, V.C., Burton-Jones, A., Sugumaran, V., Purao, S. "CONQUER: A Methodology for Context-Aware Query Processing on the World Wide Web," *Information Systems Research*, Vol. 19, No. 1, March 2008, pp. 3 – 25.

DOI: 01.IJIT.01.02.114